Twitter Connections Shaping New York City

Stanislav Sobolevsky
Center for Urban Science and Progress, New York University, Brooklyn, NY, USA
Institute Of Design And Urbanism of University Of Information Technology, Mechanics And Optics (ITMO), Saint-Petersburg, Russia
sobolevsky@nyu.edu

Philipp Kats
Center for Urban Science and Progress, New York University, Brooklyn, NY, USA
philippk@nyu.edu

Sergey Malinchik
Lockheed Martin Advanced Technology Laboratories, Cherry Hill, NJ, USA
sergey.b.malinchik@lmco.com

Mark Hoffman
Lockheed Martin Advanced Technology Laboratories, Cherry Hill, NJ, USA
mark.hoffman@lmco.com

Brian Kettler
Lockheed Martin Advanced Technology Laboratories, Cherry Hill, NJ, USA
brian.p.kettler@lmco.com

Constantine Kontokosta
Center for Urban Science and Progress, New York University, Brooklyn, NY, USA
Dept. of Civil & Urban Engineering, New York University, Brooklyn, NY, USA
ckontokosta@nyu.edu

Abstract

Geo-tagged Twitter has been proven to be a useful proxy for urban mobility, this way helping to understand the structure of the city and the shape of its local neighborhoods. In the present work we approach this problem from another angle by leveraging additional information on Twitter customers mentioning each other, which might partially reveal their social relations. We propose a novel way of constructing a spatial social network based on such data, analyze its structure and evaluate its utility for delineating urban neighborhoods. This delineation happens to have substantial similarity to the earlier one based on the user mobility network. It leads to an assumption that the social connectivity between the users is strongly related with the similarity in their mobility patterns. We justify this hypothesis enabling extrapolation of the available user mobility patterns as a proxy for social connectivity and building a network of hidden ties based on the mobility pattern similarity. Finally, we evaluate the socio-economic characteristics of the partitions for all three networks of all mentioning, reciprocal mentioning and the hidden ties.

1. Introduction

Recently, the datasets on human mobility and interactions saw increasing number of urban applications. Cell phone connections [1, 5-6, 9, 19-22], credit card transactions [24-26], GPS readings [27] as well as various sensors data [12, 28] serve as a useful proxy for human mobility, however its availability is limited largely due to the privacy concerns [2, 4, 11]. Social media, such as Twitter, on the other hand provides a broadly and more easily available alternative which once geo-tagged has been also proven useful for the human mobility studies [7, 10, 17-18], despite its limited representativeness, being often seen as a challenge.

However, besides the spatial information and the contents of the tweets, Twitter data also highlights

URI: http://hdl.handle.net/10125/50014
ISBN: 978-0-9981331-1-9

possible connections between the users, when they mention each other. While the very fact of one user being mentioned by the other does not necessarily evidence the connection between them, as people often mention influencers who certainly might not know all the persons mentioning them, mentioning data might still point out the social links, especially when reciprocal mentioning is considered.

The network of mentioning between Twitter users might be considered as a proxy for a city's social network with respect to its possible representativity bias. As the spatial projection of the social network structure is known to often reflect useful geographical information, allowing delineating regions at country [1, 20-21], global [7] or local urban scales [8, 13], we will evaluate the utility of the new mentioning network from that perspective.

A certain methodological challenge in constructing this network in space is the uncertainty about the residence location of the users. We know their locations during the activity but have no information on their actual residence or any other primary attachment. This is actually a common issue for many similar datasets – in most cases user home locations are either not known or not available due to privacy concerns. Usually in such cases researchers try to infer the most likely home location for each individual and there is a handful of approaches available for that [3]. However, this way one has to discard a lot of potentially valuable information. First, majority of the users have low activity and have to be discarded due to the lack of reliability in their home location definition. Second, even for the remaining users, selecting one single location to attach them to discards information about places they visit, while the suggested home location might still be inaccurate. In the present paper we propose an alternative approach for constructing a spatial social network based on the available geo-tagged social connections that can be used not only for the Twitter data but also for other similar datasets, such as cell-phone call records.

The resulting social network including spatial information on the users (i.e. spatial network) will be then evaluated by applying it to the neighborhood delineation in New York City. While resulting neighborhoods seem meaningful, their similarity to the neighborhoods obtained from the user mobility network analysis [18] give rise to another research question – to what extent social connections and mobility are related?

We will address this question by establishing a relation between the similarity of user mobility patterns and the chance that there exists a possible social connection reflected by a reciprocal mentioning relation between those users. This will open up potential for inferring social relations based on mobility information that might sometimes be available, while social relations are not (e.g. GPS data or credit card transactions of the customers).

2. The dataset

Twitter is a popular micro-blogging service that allows people to post short messages and follow other people across the world. In the first quarter of 2016, number of its monthly active users worldwide exceeded 310 million. Due to its (recently removed) text size limitation, users tend to generate posts on a fast pace, through multiple native and third-party applications. Due to platform popularity, any approach based on its data is a-priory applicable to the most urban areas across the globe. With its large collection of historical records, and detailed information about time, user, application, post geographical coordinates and the body of message, Twitter has a premise to be a source of abundant information on characteristic of urban landscape.

A feed of tweets with geo-locations from 5 boroughs of New York City was collected for two years, namely 2015 and 2016, using official API. Data contains the content of the tweet, id of the user and the location associated with the tweet. Tweets considered as automated were removed from consideration, as we want to focus on individual activity. The structure of the data is illustrated in the table 1 below.

Field	Meaning
Timestamp	Time of the tweet
ID	Unique ID of the tweet
UserID	ID of the user who created the tweet
Content	Content of the tweet
Hashtags	Hashtags used
Mentions	Users mentioned
Lon	Longitude
Lat	Latitude

Table 1. The structure of the Twitter data.

For the purpose of the further analysis the data has been then aggregated to 2166 census tracts across the city by mapping lon and lat fields into them.

Our final database contains over 10 million tweets from about 1,300,000 unique users. Out of those over 115,000 users were seen in at least 10 different census tracts, so provide a mobility pattern detailed enough for the future analysis.

Besides content and location, the data contain the information if the tweet mentions any other user. This way about 420,000 users were mentioning or have been mentioned by someone. But those mentions are directed and when user B is mentioned by A, user A might be never mentioned by B (this situation might not represent any actual social connection). Out of those, around 1.1M mentions where reciprocal (i.e. user A mentions B and B mentions A) involving the total of 72,500 users.

3. Constructing the Spatial Network of Mentioning

Even though not all the mentions actually represent social relations, many of them might, especially the reciprocal ones. Those relations might provide useful information on how people from different places around the city are connected with each other, and reflect the social structure of the city. For the least we construct the spatial network of mentioning, where locations around the city (census tracts) are represented with network nodes, while connections between them are represented with the network edges weighted by the total number of times users from one location mention the users from the other. This way the network is directed. We also consider its version called reciprocal mentioning network, where only reciprocal relations between the users are taken into account (the network is still directed as the number of connections in both directions between the users having a reciprocal relation might still be asymmetric).

The major challenge in constructing such a network is uncertainty of the user location. A common approach in such circumstances is inferring the most likely residential location for each customer based on his/her mobility pattern [3]. However, this is only possible for the most active customers, leaving us with uncertainty for the low-active ones, which actually represent majority of the users to be excluded from the further consideration together with all their connections. Still the residential locations for the remaining ones might not always be correct. Also many users have multiple centers of activity basically belonging to various local communities. Thus, attaching the users to just one of those centers ignoring the rest might not always make sense.

Instead of having to deal with this uncertainty and having to filter out substantial part of the available data, in the present work we propose an alternative approach. Instead of defining one single home census tract for each user we consider uncertain attachment of the user to different locations visited with the probabilities proportional to the intensity of the visits. This way each user will be taken into account and his/her mentioning activity will be distributed among all the visited census tracts proportionally to the frequency of the visits. Mathematically this can be represented as:

$$M(A,B) = \sum_{u \neq v} m(u,v) \frac{V(u,A)}{\sum_C V(u,C)} \frac{V(v,B)}{\sum_C V(v,C)}$$

where M(A,B) is the network edge weight between the nodes (census tracts) A and B, m(u,v) is the number of times user u mentions user v, V(u,A) is the number of times user u tweets from census tract A. The reciprocal version is:

$$MR(A,B) = \sum_{u \neq v, m(u,v) m(v,u) \neq 0} m(u,v) \frac{V(u,A)}{\sum_C V(u,C)} \frac{V(v,B)}{\sum_C V(v,C)}$$

Analyzing the content of the activity might be further useful to better understand the nature of social connections seen in Twitter (e.g. positive, negative or neutral context in which users mention each other). This could be subject of further study.

4. Neighborhood delineation

Networks of human mobility and social networks often reflect the geographic structure of the area at a regional [1, 20-21] or even global scale [7]. This has been also validated on the city scale by using cell phone and taxi data [8]. Mobility patterns extracted from Twitter data have been successfully utilized to discover neighborhoods of New York City (NYC) [18]. This work considers the network of locations across the city from another perspective, being connected whenever a user residing in one location performs activity in the other and compares this network and its delineation results against the commuting network based on the Longitudinal Employment Household Dynamics Data from US Census and Twitter networks.

In the present work we apply the same partitioning algorithm Combo [23] to the above networks of Twitter mentioning and reciprocal mentioning. The algorithm optimizes the partition quality quantified based on the modularity function [14] using a combination of splits, joins and merges

being iteratively applied to any initially chosen partition. The partition results for all mentioning and reciprocal mentioning networks (those two are chosen to compare the pattern created by all possible links including those when one user knows the other but not the other way around and only the strong links where users know each other) are reported on the Figs. 1 and 2. They confirm the previously noticed pattern that communities of the networks of human mobility and interactions use to be spatially cohesive [1, 20-21]. They reveal key areas of NYC and are to a large extent consistent with the previous findings of [18], considering the network of locations across the city from a different perspective as described above. Reciprocal network actually provides a stronger similarity also capturing important features such as airports being attached to the core business area (Manhattan and Downtown Brooklyn).

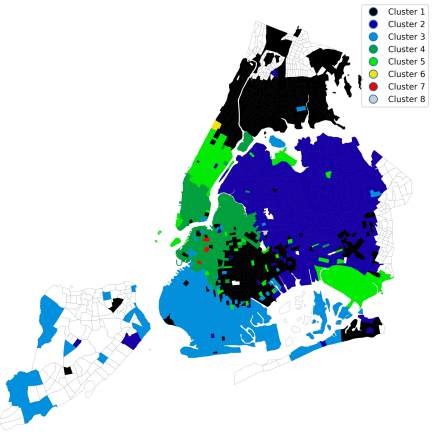

Figure 1. Partition of NYC based on all mentioning relations. Same/different colors show areas belonging to the same/different communities.

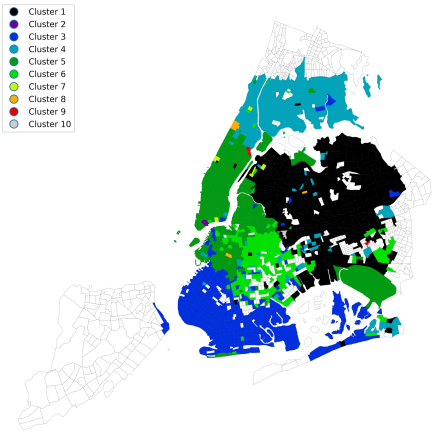

Figure 2. Partition of NYC based on reciprocal mentioning relations. Same/different colors show areas belonging to the same/different communities.

5. Mobility patterns vs mentioning relations

The similarity between the structure of Twitter mentioning and mobility networks gives rise to an important hypothesis that the social connectivity between the users is strongly related to the similarity of their mobility patterns and that the least could be used as a proxy to the first. Previously it has been already noticed that people who are connected use to meet each other first [15]. Now we aim to confirm that the similarity in mobility patterns could serve as a quantitative proxy to the connectivity.

This could be important for example when mobility information is available while social connections are not, like in case of GPS readings or credit card transaction data. Even in case of Twitter not all the social relations existing between the users are reflected by their mentioning – many more relations might exist, but stay hidden from our attention as those users might never mention each other on Twitter. While if similarity of user mobility patterns is indeed related to the social connectivity, then those hidden relations might be inferred based on the extensive mobility information contained in the geo-tags of the tweets.

In order to evaluate the hypothesis, we consider the average chance for a pair of users to be connected

by reciprocal mentioning as a function of the cosine similarity of their mobility patterns (CSMP). Cosine similarity is defined as

$$CSMP(u,v) = \frac{\sum_A V(u,A)V(v,A)}{\sqrt{\sum_A V(u,A)^2 \sum_A V(v,A)^2}}$$

where mobility pattern of a given user u is simply determined by the number of times $V(u,A)$ user u tweeted from each census tract A. The figure 3 shows a relation between the level of CSMP score and the frequency of reciprocal mentioning connections between the pairs of users having this given similarity score between their mobility patterns. This relation looks like a steady and nearly linear increase, meaning that the chance for a pair of users to be connected is nearly proportional to the similarity of their mobility patterns – it is much more likely for the users who visit the same places to be connected than for the users whose mobility patterns barely overlap. Of course even for the users with nearly identical mobility patterns the chance of being connected is far from 100% (close to 1% only) - partly because not all the connections are reflected in the mentioning data, partly as in a huge city a pair of users can be accidentally captured by Twitter data in the nearby places but never get to know each other. Nevertheless, a nearly linear relation allows suggesting the CSMP as a proxy for the social connectivity as long as we are going to use it at the aggregated scale and care more about the relative magnitude rather than about the exact value of the number of social connections between the two locations. While for each specific pair of individuals it is of course not possible to make any reliable conclusion on whether they are connected or not based on their mobility patterns alone, the nearly linear relation on figure 3 shows that the cumulative CSMP score provides an estimate for the average connection frequency, which can be efficiently used as a proxy for the number of actual connections at the aggregated scale for a large enough group of users.

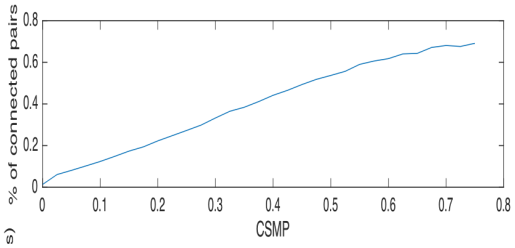

Figure 3. Relation between the cosine similarity of mobility patterns (CSMP) for pairs of users and their chance of getting connected

6. Network of hidden ties

As the Twitter mentioning is likely to reflect only a small portion of existing social relations between the Twitter users, the mentioning network might not be a comprehensive proxy to the actual social network. However, one can construct the network based on the anticipated relations (hidden ties) using CSMP as a proxy. Specifically, construct a network of census tracts where each edge between a pair of tracts A and B is weighted by the projected number of reciprocal connections between users from A and B based on the cumulative CSMP score between all pairs of users from A and B. Like before, the users are attached to the locations based on the approach from section 3.

The partitioning of the network of mobility pair similarities, which one can also call a network of hidden connections, is presented on the figure 4. It provides a very clear delineation of the core business area (Lower Manhattan and Downtown Brooklyn), Western and Eastern parts of Upper Manhattan, and the residential areas of Bronx, Brooklyn and Queens as one single community. However, it also captures another important pattern missed by the other networks– link between Staten Island and Battery Park area in Manhattan where Staten Island ferry departs. Also worth mentioning that now we're able to include many more areas that were otherwise skipped due to the sparseness of the mentioning data.

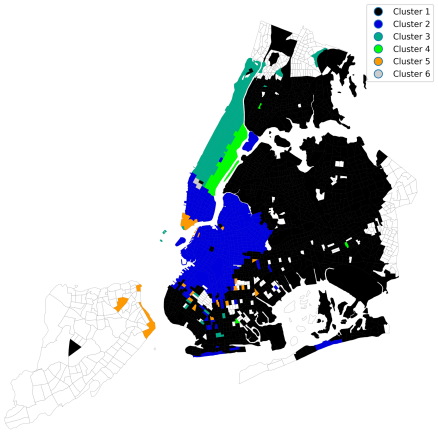

Figure 4. Partitioning of the network of hidden ties. Same/different colors show areas belonging to the same/different communities.

7. Socio-economic properties of the partitions

This section provides a quantitative interpretation of the partitioning in terms of their socio-economic profile. As all three partitions have distinctive spatial representation, comparing the social and economic properties of the territories they cover might provide the socio-economic context for the obtained communities of people. In order to do so, we collected a list of properties using U.S. Census 2015 American Community Survey (ACS)[1]. Normalized average characteristics of each community from each partition for all three considered networks are presented in the Figures 5-7. Fig. 8 represents distribution of the characteristics over communities for each partition (positive values mean above city average, negative – below).

Most of the communities seem to have distinctive socio-economic profile and shaped by their properties, such as median income, commute time, population density, and others. E.g. all three network distinguish 1-2 dense wealthy neighborhoods with a low average commute time. While networks of mentioning also provide a good distinction by the age of the customers, the network of hidden links fails to do so. Table 2 provides a quantitative characteristic of how distinctive the socio-economic profiles of different partitions are or equivalently – how homogeneous the characteristics inside the communities are. This is done by measuring the fraction of variance of each parameter realized inside the communities

$$H = \frac{\sum_A (x_A - E[X_{c(A)}])^2}{\sum_A (x_A - E[X])^2}$$

where A runs through all census tracts, $c(A)$ is the community of the tract A in a given partition, x_A is a considered feature value for the census tract A, X is a distribution of the feature values over the city, while X_c is a distribution over the community c, E stands for the mean. The H is a normalized metric between 0 and 1 and the lower it is, the more homogenous is the given characteristic inside the communities and the more distinctive those communities are from one another.

Different partitions work differently in terms of splitting the city by each socio-economic characteristic. For example, partition of the hidden ties network provides the best separation in terms of the population density, the partition of the reciprocal connections network works the best in terms of age, average commute time and the percent of homeowners, while partition of all connections network works the best for the median income. Overall, the partition of the network of reciprocal ties provides the best socio-economic separation of the city. However lower overall performance of the network of hidden ties could be explained by having a smaller number of communities it produces (clearly, the more communities the partition has, the more socio-economic differences could be captured). This could be overcome by controlling for the number of communities produced – if a more fine-grained partition is needed one may introduce a resolution parameter into community detection [16].

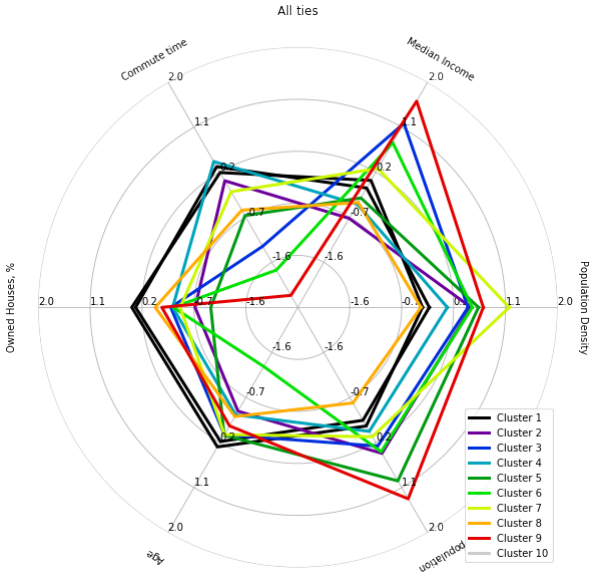

Figure 5. Average socio-economic properties of the communities of all mention network partitions

[1] https://www.census.gov/programs-surveys/acs/

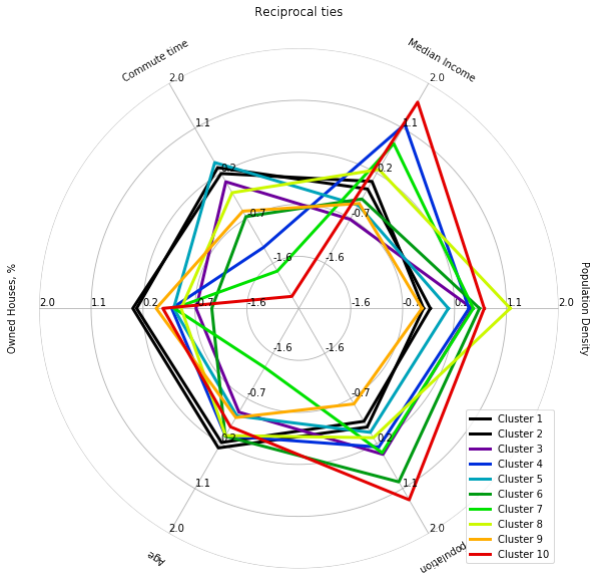

Figure 6. Average socio-economic properties of the communities of the reciprocal mention network

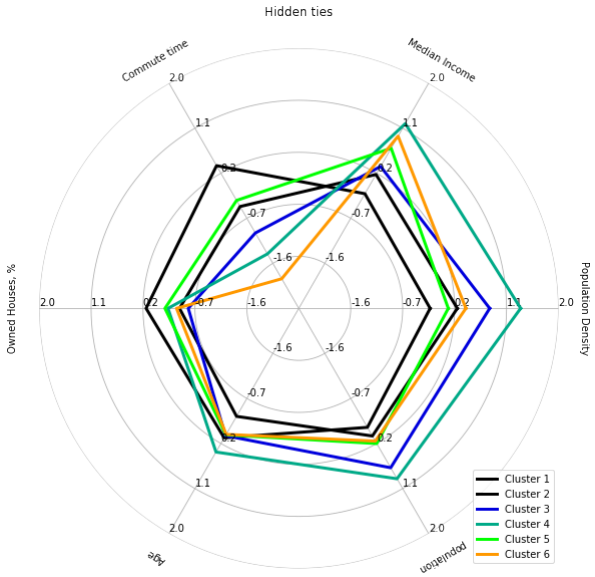

Figure 7. Average socio-economic properties of the communities of the hidden ties network

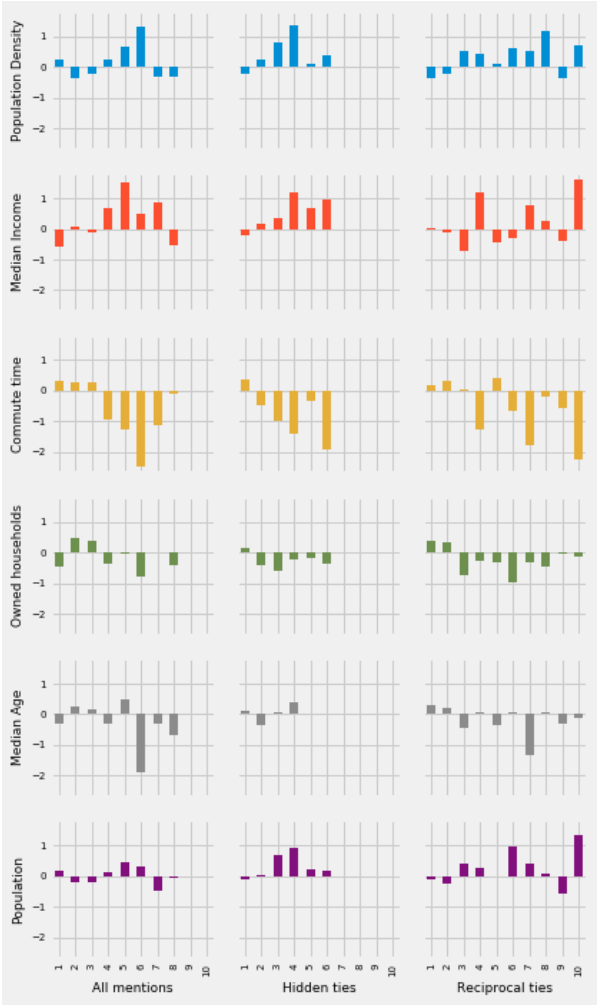

Figure 8. Distribution of normalized socio-economic features over partitions

	All mentions	Reciprocal ties	Hidden ties
Age	0.91	0.90	0.94
Median Income	0.67	0.68	0.92
Population Density	0.89	0.90	0.86
Average Commute	0.69	0.60	0.69
% of Homeowners	0.82	0.74	0.84
Average per partition	0.80	0.76	0.85

Table 2. Variance Ratio of the socio-economic parameters distribution within each partition.

Conclusions

We presented a novel approach for constructing spatial networks of interactions between users with uncertain residence locations and applied it to constructing the networks of all and reciprocal mentioning between NYC Twitter users. The structure of those networks turned out to be useful for delineating major neighborhoods across the city, especially for the network of reciprocal mentioning. The similarity of this structure to the earlier studied network of users' mobility gives rise to the hypothesis that social links between people are related to the similarity between their mobility patterns. We validate this hypothesis by showing that the chance for a pair of people to be connected is in nearly linear relation to the cosine similarity of their mobility patterns (CSMP). Based on this finding we construct the network of anticipated hidden ties using the CSMP scores as a proxy. This network is seen to provide additional useful insights on the neighborhood structure of NYC, emphasizing the utility of the CSMP as a proxy for the social connectivity, especially in cases when ground-truth information on social connectivity is missing or sparse. In conclusion, the comparative socio-economic analysis of the resulting partitions of all three networks is provided, showing that the resulting communities are distinctive in their socio-economic characteristics.

References

1. Amini Alexander, Kevin Kung, Chaogui Kang, Sobolevsky S., and Carlo Ratti. The impact of social segregation on human mobility in developing and industrialized regions. EPJ Data Science, 3(1):6, 2014
2. Belanger France, R. E. Crossler. Privacy in the digital age: a review of information privacy research in information systems. Mis Quarterly, 35:1017–1042, 2011
3. Bojic Iva, Massaro, E., Belyi, A., Sobolevsky, S., & Ratti, C. (2015, December). Choosing the Right Home Location Definition Method for the Given Dataset. In SocInfo (pp. 194-208).
4. Christin Delphine, A. Reinhardt, S. S. Kanhere, and M. Hollick. A survey on privacy in mobile participatory sensing applications. Journal of Systems and Software,84:1928–1946, 2011
5. Girardin Fabien, Francesco Calabrese, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. Digital foot printing: Uncovering tourists with user-generated content. Pervasive Computing, IEEE, 7:5276, 2008
6. Gonzalez M. C, C.A. Hidalgo, and A.-L. Barabasi Lazlo, Understanding individual human mobility patterns. Nature, 453:779–782, 2008
7. Hawelka Bartosz, Izabela Sitko, Euro Beinat, Sobolevsky S., Pavlos Kazakopoulos, and Carlo Ratti. Geo-located twitter as proxy for global mobility pattern. Cartography and Geographic Information Science, 41(3):260–271, 2014
8. Kang, Chaogui, Sobolevsky, S., Liu, Y. and Ratti, C., Exploring human movements in Singapore: a comparative analysis based on mobile phone and taxicab usages. In Proceedings of the 2nd ACM SIGKDD international workshop on urban computing (p. 1). ACM. (2013)
9. Kung Kevin, Kael Greco, Sobolevsky S., and Carlo Ratti. Exploring universal patterns in human home/work commuting from mobile phone data. PLoS ONE, 9(6):e96180, 2014
10. Kurkcu, Abdullah, Ozbay, K., & Morgul, E. F. (2016). Evaluating the Usability of Geo-located Twitter as a Tool for Human Activity and Mobility Patterns: A Case Study for New York City. Proceedings of the 95th TRB Annual Conference, #16-3901, Washington, D.C., January, 2016
11. Lane Julia, V. Stodden, S. Bender, and H. Nissenbaum. Privacy, big data, and the public good. Cambridge University Press, 2014
12. Lathia Neal, Daniele Quercia, and Jon Crowcroft. The hidden image of the city: Sensing community well-being from urban mobility. In Judy Kay, Paul Lukowicz, Hideyuki Tokuda, Patrick Olivier, and Antonio Kruger, editors, PervasiveComputing, volume 7319 ofLecture Notes in Computer Science, pages 91–98.2012
13. Louail Thomas, Maxime Lenormand, Oliva Garcia Cantu Ros, Migueal Picornell, Ricardo Herranz, Enrique Frias-Martinez, Jose J. Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. Scientific Reports, 4:5276, 2014
14. Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), 8577-
15. Calabrese, F., Smoreda, Z., Blondel, V.D. and Ratti, C., Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. PloS one, 6(7), p.e
16. Reichardt, J. and Bornholdt, S., Statistical mechanics of community detection. Physical Review E, 74(1), p.016110.
17. Paldino Silvia, Iva Bojic, Sobolevsky S., Carlo Ratti, and Marta CGonzalez. Urban magnetism through the lens of geo-tagged photography. EPJ Data Science, 4(1):1–17, 2015
18. Qian, Cheng., Kats, P., Malinchik, S., Hoffman, M., Kettler, B., Kontokosta, C. and Sobolevsky, S., 2017, July. Geo-Tagged Social Media Data as a Proxy for Urban Mobility. In International

Conference on Applied Human Factors and Ergonomics (pp. 29-40). Springer, Cham.
19. Quercia Daniele, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone location data. In Data Mining(ICDM), 2010 IEEE 10th International Conference on, pages 971 –976, 2010
20. Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R. and Strogatz, S.H., Redrawing the map of Great Britain from a network of human interactions. PloS one, 5(12), p.e
21. Sobolevsky S., Michael Szell, Riccardo Campari, Thomas Couronne, Zbigniew Smoreda, and Carlo Ratti. Delineating geographical regions with networks of human interactions in an extensive set of countries. PloS ONE,8(12):e81707, 2013
22. Reades Jonathan, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. Cellular census: Explorations in urban data collection. Pervasive Computing, IEEE, 6:30–38, 2007
23. Sobolevsky S., Campari R., Belyi A., Ratti C., A General Optimization Technique for High Quality Community Detection in Complex Networks, arXiv:1308.3508 (2013).
24. Sobolevsky S., Iva Bojic, Alexander Belyi, Izabela Sitko, Bartosz Hawelka, Juan Murillo Arias, and Carlo Ratti. Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity. In 2015 IEEE International Congress on Big Data, pages 600–IEEE, 2015
25. Sobolevsky, S., Sitko, I., Des Combes, R.T., Hawelka, B., Arias, J.M. and Ratti, C., 2014, June. Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain. In Big Data (BigData Congress), 2014 IEEE International Congress on (pp. 136-143)
26. Sobolevsky, S., Sitko, I., des Combes, R.T., Hawelka, B., Arias, J.M. and Ratti, C., Cities through the prism of people's spending behavior. PloS one, 11(2), p.e
27. Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S.H. and Ratti, C., Quantifying the benefits of vehicle pooling with shareability networks. Proceedings of the National Academy of Sciences, 111(37), pp.13290
28. Constantine E Kontokosta and Johnson N. Urban phenology: Toward a real-time census of the city using wi-fi data. Computers, Environment, and Urban Systems, 2016